# Learning and Recognizing Structures in Façade Scenes (eTRIMS) – A Retrospective

Lothar Hotz, Bernd Neumann

## 1 Introduction

"eTraining for the Interpretation of Man-made Scenes (eTRIMS)" was the official title of an EU project which ended September 2009 after a duration of 42 months. Five research teams constituted the consortium:

- Institute of Photogrammetry, University of Bonn (Wolfgang Förstner, coordinator)
- Center for Machine Perception, University of Prague (Radim Sara)
- Department of Electrical and Electronic Engineering, Imperial College, London (Maria Petrou)
- Cognitive Systems Laboratory, University of Hamburg (Bernd Neumann)
- Hamburg Informatics Technology Center, Hamburg (Lothar Hotz)

The project was allocated in the "Cognitive Systems and Robotics" unit in IST-FP6 and had the main goal to develop learning methods for models of spatial structures such as roofs in aerial images or window, balcony and door configurations of façades. The focus was on learning methods which would allow to continuously adapt and extend the model base of an image interpretation system. All partners were engaged in work on this main topic. In addition, Bonn was responsible for organising a database of example images and for evaluation. Hamburg had to develop an infrastructure for the integration of learning results with an interpretation system.

In this article, we first give an overview of the learning approaches, the image processing modules, and the interpretation framework which have been integrated in the infrastructure. We then describe two learning approaches developed in Hamburg in more detail, one of which was used for a continuous step-by-step learning process which provided interesting insights about an advantageous ordering of examples. Finally, we compare our experiences with crisp logic-based and probabilistic scene interpretation, both of which haven been realized in eTRIMS.

## 2 System Overview

We begin the system overview by describing the structure of the scene interpretation system developed for eTRIMS. Figure 1 illustrates the architecture and the main components.

The system is structured into a layer for low-level image analysis, a layer for high-level interpretation, and a middle layer mediating between the two. In addition, a user interface allows manual image selection (usually from the eTRIMS image database) and screening of annotations.

The knowledge base at the left indicates that learnt declarative models play a part at all levels. In the middle layer, a

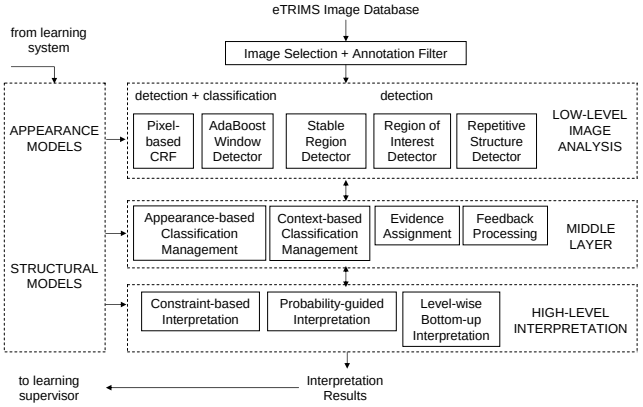


Figure 1: Structure of scene interpretation system. The boxes represent modules developed by all partners of eTRIMS. Modules can be configured in various combinations.

learnt decision tree (DT) may be employed for soft classification of regions delivered by any of the region detectors of the low-level layer. High-level interpretation can be guided by learnt structure models of various kinds, for example by a Bayesian Compositional Hierarchy (BCH) [1].

The second main part of the system infrastructure is a learning testbed, illustrated in Figure 2. The boxes in the middle row denote learning procedures to obtain appearance models for individual objects. They can be trained and tested without recourse to the interpretation system.

The boxes in the lower row denote learning procedures for structural models whose evaluation may require high-level scene interpretation. Their output can be embedded into the interpretation system as declarative high-level models. For evaluation, a Learning Supervisor has been developed which accepts the output of the interpretation system, compares it with corresponding annotations and determines the next leaning step. Thus, an automatic learning cycle has been realized.

This architecture allows for testing and evaluating various combinations of modules and enables a systematic comparison of different scene interpretation approaches (see [2]).

## 3 Structural Learning and Continuous Learning

High-level interpretation of scenes is based on a compositional hierarchy of aggregate models. For the domain of building façades, these models should describe the typical spatial con-

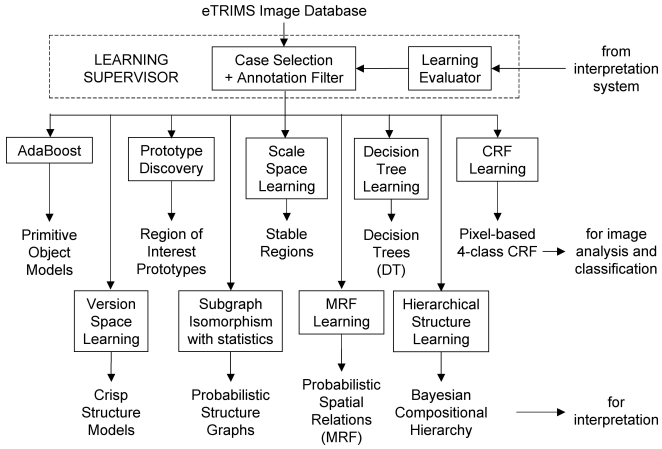


Figure 2: Structure of learning system. The boxes in the middle show learning methods for image analysis and classification. The boxes below show learning methods for high-level interpretation support.

figurations of façade objects relative to each other. One of the central goals of eTRIMS was to develop learning methods which allow to obtain such models from examples and to adapt models if new façade structures evolve.

One innovative approach developed in eTRIMS for learning aggregate models is an extension of Version Space Learning (VSL) [3] which exploits positive and negative examples (presented by a teacher) to generate crisp structural models. It employs an expressive and extensible description language for spatial properties and relations, embedded in a generalisation hierarchy.

To establish a version space, the concept description language must allow an ordering according to generality, i.e. a taxonomy of concepts. While simple conjunctive attribute languages trivially allow such ordering, more expressive description languages suitable for the façade domain have not been investigated before. In eTRIMS, such a language has been developed and successfully used for concept learning within the VSL framework.

Particular attention has been given to structural relations, described in the partonomy section of an aggregate concept. Structural relations are specified by

- the total number of aggregate parts in terms of an attribute of range type,
- the different types of each part in terms of an attribute of set type, and
- the number of parts for each type in terms of range attributes.

This way, conceptual structure descriptions can be ordered according to generality.

In a further learning approach investigated in eTRIMS, probabilistic structural descriptions for aggregates have been obtained, again in a supervised-learning setting, but using only positive examples. The learning process basically seeks for the minimal generalisation of all observed spatial structures, but also anticipates unseen structures to a certain extent. The result is an aggregate model causing significantly fewer interpretation errors than a crisp model, learnt from the same training set and evaluated on the same test set. This approach is based on a new modelling and learning method for aggregate models called Probabilistic Structure Graphs (PSGs). The main feature of PSG models is an unrestricted joint probability distribution for the existence of object parts. This is different from e.g. the Markov Random Field approach which only allows to model conditional dependencies of nodes w.r.t. to a subset of all nodes in the graph because of the Markov Property.

PSG learning makes use of two kinds of generalisations, structure generalisation and relation generalisation. Structure generalisation is performed by integrating new examples into the current model by Attributed Subgraph Isomorphism. By this operation, the structure of the current model remains the same if the new example is structurally contained in the model, and it is extended if the new example has a part not yet covered by the model.

Relational generalisation is performed using a hierarchy of spatial relations, shown in Figure 3. If corresponding spatial relations differ between the current model and a new learning example, the model relation is generalised to include the example.

Learnt aggregate concepts have been used for scene interpretation in SCENIC and have shown their potential for aggregate recognition and for hypothesising missing evidence and thus supporting high-level feedback to low-level image analysis.

One of the long term goals of research in machine learning is the development of curricula. A curriculum contains *the set of courses, course work and content offered* (Wikipedia) to a scholar. Current learning is either batch learning, i. e. all training examples are available at the time of learning, or - mostly in the context of a mobile robot - on-line learning, where the training examples appear sequentially, in the course of robot activities, resulting in an uncontrolled sequence of training examples. In contrast, a curriculum specifies the sequence learning tasks and, consequently, the sequence of course material.

When learning structural models with VSL, the effect of different orderings of training examples on the learning rate was investigated in eTRIMS. The learning rate was determined as the number of recognition failures remaining after a certain number of learning examples. In contrast to what one might expect, learning appeared to be more successful if complex examples, spanning the space of structural variations, were presented first.

# 4 Scene Interpretation in eTRIMS

The scene interpretation system SCENIC, developed in Hamburg, has the following characteristics:

- Scene interpretation transforms primitive objects into higher-level meaningful units.
- Scene interpretation is based on declarative conceptual models.
- Conceptual models are organized as aggregates in a compositional hierarchy.
- Concepts distinguish between 3D models and their views.
- Scene interpretation is a stepwise process.

These features are largely undisputed in the scene interpretation community [4, 5, 6], but a few remarks concerning the realization in SCENIC are in order. First, interpretation is not considered as a strictly symbolic process. Primitive objects may have quantitative properties (such as location and shape) and

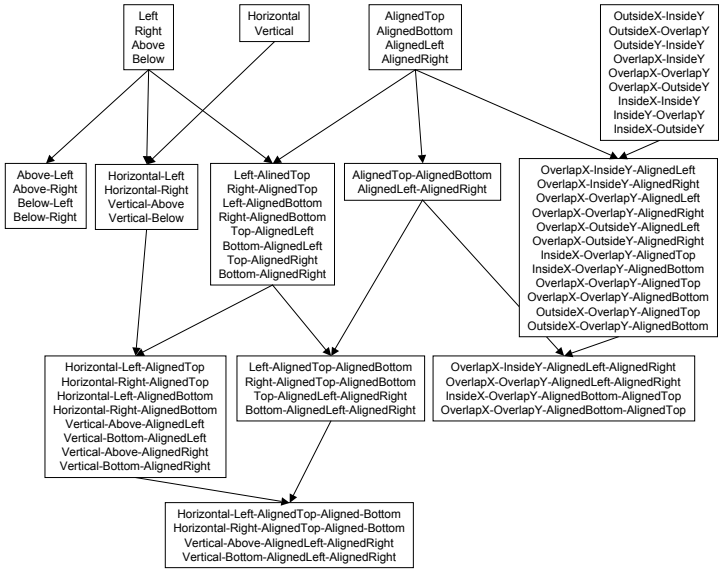


Figure 3: Hierarchy of spatial relations used by Probabilistic Structure Graph models

may be related to each other by quantitative relations (such as distance and orientation). Second, the compositional hierarchy is not a partonomy in any strict mereological sense. An aggregate (Figure 4) simply a named set of constituents satisfying certain conditions. For example, a wall, railing, window, and door in a certain spatial arrangement constitute a balcony. A third remark concerns the stepwise process, which may not seem compelling for the interpretation of a static scene. We consider stepwise interpretation as a general framework for scene interpretation because of the possibility to guide interpretation steps by an evolving context. This is, of course, indispensable for real-time interpretation of time-varying scenes.

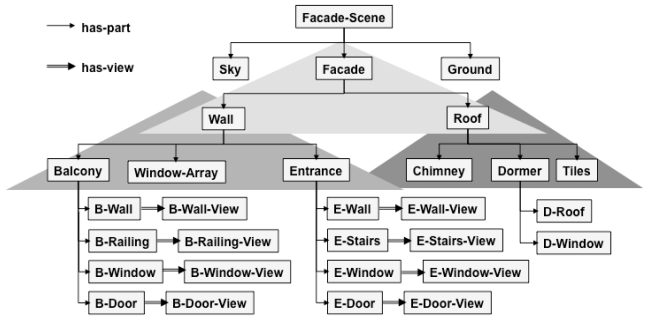


Figure 4: Aggregate structure of compositional hierarchy of façade scenes.

In the course of the project, the original logic-based ("crisp") interpretation system SCENIC mutated to a probabilistic interpretation system, and it is interesting to review this development. In crisp SCENIC, stepwise interpretation was modelled as a search for a partial "logical model" of the conceptual knowledge base, conforming to the formal view of scene interpretation as logical model construction [7]. All evidence must be classified as instances of concepts, higher-level instances are formed as required by the compositional hierarchy, and missing evidence is hypothesized. The conditions for parts of an aggregate were expressed as crisp constraints (value ranges or possible choices) and evaluated by a powerful constraint system. Interpretation steps could be viewed as bottom-up or top-down steps navigating in compositional and taxonomical hierarchies [8].

The main motivation for introducing probabilistic representations was to provide a preference measure for logically ambiguous interpretation steps. For example, a window can be assigned to an entrance aggregate, a balcony, or a window array, and such decisions need guidance to decrease the chance of logical inconsistencies later in the interpretation process. To this end, Bayesian Compositional Hierarchies (BCHs) were developed which replaced the crisp constraints in aggregate concepts by joint probability distributions (JPDs) [1] and allowed to compute the probabilities for competing partial scene interpretations. Different from a general Bayesian Network, a BCH has a probabilistic dependency structure homomorphous with the structure of the compositional hierarchy, but with arbitrary dependencies within each aggregate. A drawback was, however, that alternative aggregate structures had to be modelled in alternative BCHs, giving rise to hundreds of alternative façade models.

In this probabilistic setting, stepwise interpretation decisions are essentially realized as evidence assignment for primitive objects of a compositional hierarchy, in parallel for alternative models. After each assignment, evidence is progated to the rest of the façade model, providing new prior probabilities for further assigments. This way, a preference measure is available to perform beam search within a limited set of most probable alternatives. Finally, higher-level interpretations are obtained as maximally probable values for hidden aggregate variables. In experiments, it could be shown that the computation of dynamic priors improves the evidence classification performance [9].

Probabilistic propagation can be compared to hypothesis-generation steps in logic-based interpretation. Both exploit the

evolving high-level context, but propagation has the advantages (i) to generate expectations for complete scenes (given corresponding models) and (ii) to allow updates without backtracking.

# 5 Summary

The aim of the project eTRIMS has been to advance the state of the art of cognitive systems by developing a methodology for autonomous and continuous learning. The project has concentrated on structural learning, where spatial relations between components and compositional hierarchies play a central role. Such learning is particularly relevant for the interpretation of man-made objects, hence, the project has used the recognition of buildings and parts of buildings in outdoor scenes as its exemplary application domain. Due to the diversity of shapes and spatial arrangements of the different parts of a building, any such recognition system must be capable of continually updating its conceptual knowledge. This was the motivation for the development of innovative methods for continuous learning.

Published deliverables and articles can be found at the eTRIMS website *www.ipb.uni-bonn.de/projects/etrims*. The website also contains benchmarks that come with ground truth data and evaluation criteria for enforcing competitions in the area of scene interpretation.

# Acknowledgments

This research has been supported by the European Community under the grant IST 027113, eTRIMS - eTraining for Interpreting Images of Man-Made Scenes.

# References

[1] Neumann, B.: Bayesian compositional hierarchies - a probabilistic structure for scene interpretation. Technical Report Memorandum FBI-HH-B-282/08, Department of Informatics, Hamburg University, University of Hamburg (2008)

[2] Šára, R.: D4.4: Performance evaluation (2009) eTRIMS project deliverable.

[3] Mitchell, T.: Version Spaces: An Approach to Concept Learning. PhD thesis, Stanford University, Cambridge, MA (1978)

[4] Nagel, H.H.: From image sequences towards conceptual descriptions. Image Vision Comput. **6**(2) (1988) 59–74

[5] Georis, B., Mazière, M., Brémond, F., Thonnat, M.: Evaluation and Knowledge Representation Formalisms to Improve Video Understanding. In: Proc. of IEEE International Conference on Computer Vision Systems ICVS06. (2006) 27–27

[6] Heintz, F., Doherty, P.: DyKnow: A framework for processing dynamic knowledge and object structures in autonomous systems. In: Proceedings of the International Workshop on Monitoring, Security, and Rescue Techniques in Multiagent Systems (MSRAS). (2004)

[7] Schröder, C., Neumann, B.: On the Logics of Image Interpretation: Model-Construction in a Formal Knowledge-Representation Framework. Proc. ICIP-96, Int. Conf. on Image Processing **2** (1996) 785–788

[8] Hotz, L., Neumann, B., Terzic, K.: High-level Expectations for Low-level Image Processing. In: KI 2008: Advances in Artificial Intelligence. Volume LNCS 5243., Springer (2008) 87–94

[9] Kreutzmann, A., Terzic, K., Neumann, B.: Context-aware Classification for Incremental Scene Interpretation. Proc. Workshop on Use of Context in Vision Processing (UCVP 2009) (2009)

## Contact

Dr. Lothar Hotz, Prof. Ph.D. Bernd Neumann
HITeC c/o Fachbereich Informatik, Universität Hamburg
22527 Hamburg, Germany
Tel.: +49 (0)40 42883-2451
Fax: +49 (0)40 42883-2572
Email: {neumann,hotz}@informatik.uni-hamburg.de

Bild

**Lothar Hotz** is a senior researcher at the Hamburgs Informatics Technology Center (HITeC e.V.) located at the University of Hamburg. He has participated in several projects related to topics of configuration, knowledge representation, constraints, diagnosis, scene interpretation, requirements engineering, parallel processing, and object-oriented programming languages.

Bild

**Bernd Neumann** studied Electrical Engineering at Darmstadt/FRG (Diploma 1967) and Information Theory at MIT/USA (M.S. 1968, Ph.D. 1971). He is Professor at the Department Informatik at the University of Hamburg, leading the Cognitive Systems research group and the Artificial Intelligence Laboratory in Hamburgs Informatics Technology Center HITeC. His scholarly work pertains to several AI subfields including high-level image interpretation, configuration and diagnosis. Bernd Neumann is Fellow of ECCAI and DFKI.